ARSynth : Robust Real-Time Human Torso Tracking from Synthetically Trained Deep Neural Networks

Prashanth CHANDRAN[1], Endri DIBRA[2], Ben HUBER[2]
[1] Computer Graphics Lab, ETH Zurich, Zurich, Switzerland;
[2] Arbrea Labs, Zurich, Switzerland

https://doi.org/10.15221/19.043

Abstract

Robust real-time tracking of the human body is crucial to applications that benefit from live visualizations performed on the underlying body. Such applications could fall in the category of Augmented Reality for Human Bodies, finding great usage in the broader fields of Medicine and Apparel. Specifically, robust real time tracking of the female torso is a crucial component in the pre-visualization of cosmetic breast surgeries. In order to track a torso from monocular RGB input, landmarks that describe the pose and shape of the torso have to be detected. Existing state of the art in algorithms for human pose estimation are dominated by deep neural networks and rely on the availability of large databases with high quality annotations. However, for the requirement of pre-visualizing cosmetic breast surgeries, existing databases fall short as they contain no or very few landmarks that can reliably help estimate the shape of the female torso. Therefore, by building on top of openly available databases of human character models, we create a pipeline for generating synthetic female torsos in both naked and clothed scenarios. We show that deep landmark detectors trained using such a synthetic dataset are capable of generalizing well to unconstrained real world images.

Keywords: 3D body scanning, Augmented Reality, Pose Estimation, Cosmetic Breast Surgery, Pre-visualization

1. Introduction

Cosmetic surgeries are an important use case for Augmented Reality in medicine, as they require the patient and the doctor to be on the same page with regard to their expectations for the surgery. In this work, we address problems related to cosmetic breast surgeries, where pre-visualization will tremendously help patients in making decisions about their surgical requirements (see Figure 1). Specifically, we address the problem of landmark detection on a female torso, which is a critical first step for such a pre-visualization. We create the first high quality synthetic dataset designed for cosmetic surgeries on the female torso and combine it with the power of deep learning to create a deep, data-driven landmark detector.

The rest of this paper is organized as follows. In section 2, we go over related methods in landmark detection. In section 3, we describe in detail our landmark detection architecture. Our synthetic data generation pipeline is described in section 4. The details of our implementation are provided in section 5. Qualitative and quantitative results are provided in section 6, and section 7 summarises our work.

2. Related Work

In recent years, deep learning methods have significantly advanced the state of the art in landmark detection. For a concise summary, we differentiate these methods based on their architecture and their approach to the problem. With regards to network architecture, existing work can be broadly classified into three categories namely i) networks that are a combination of convolutional and fully connected or 'dense' layers ii) fully convolutional networks, and iii) recurrent networks. The former consist of architectures that take an image as input and learn convolutional filters that extract low level and semantic features, which are then flattened and passed onto one or more fully connected or 'dense' layers. The final layer of such an architecture outputs a vector of landmark positions. On the other hand, fully convolutional architectures predict the positions of landmarks as heatmaps that encode the probability of a landmark being present at a particular pixel. Such architectures have a few advantages viz. (i) fully convolutional networks are translation invariant (ii) images of different sizes can be used at training and test times (iii) as landmarks are extracted from the heatmaps, they provide a guarantee that the predicted landmarks always lie within the domain of the image (iv) the representation of landmarks as a heatmap makes the prediction of such networks human interpretable, allowing a better understanding of why things fail. Finally, when working with a temporal sequence of images at test time, it is often necessary that some smoothing is performed on the predicted landmarks as a post processing step to ensure temporal consistency of landmarks. To incorporate this into the learning pipeline, neural network architectures where predictions from a previous frame could also be fed as input to the network have also been proposed [6].

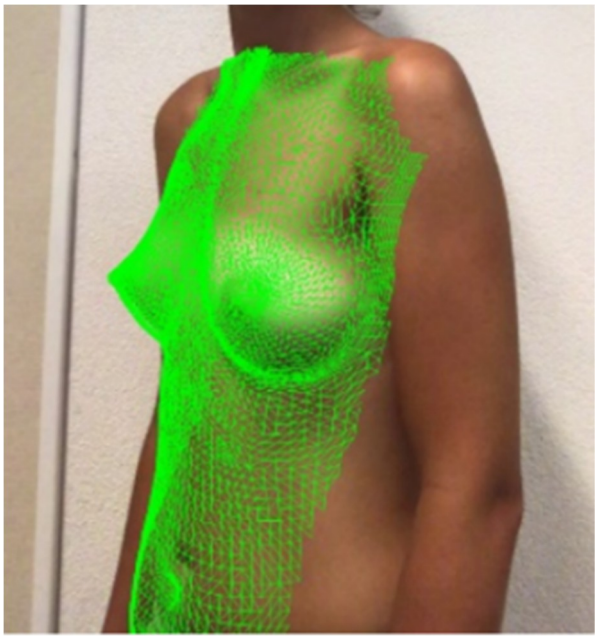

Figure 1: A real time AR torso visualization for a patient which uses the landmark detector described in this work.

Based on their approach towards solving landmark detection, architectures can be classified broadly into i) model based fitting methods ii) multi-task learning, and iii) cascaded or regional models. Model based methods assume an underlying low resolution 3D model that is parametrically fit to images using learned features. Multi-task methods following the principle of 'auxiliary learning' jointly infer multiple attributes of the given image. For example, in the case of facial landmark detection, jointly determining additional attributes of the image such as the person's age, gender, head pose etc have been shown to improve the accuracy of landmark detection [7].
Since an exhaustive summary of all landmark detectors is beyond the scope of this work, we refer the readers to [1, 2, 3, 4, 5, 6, 7] for a summary of recent methods that we reviewed to develop our solution.

3. Methodology

Our cosmetic breast surgery pre-visualization pipeline involves a number of stages. It begins with the acquisition of images of the patient's torso. Landmarks are detected on the acquired imagery. A parametric model of the torso is fit to these landmarks and additional visual cues from the image to match the shape and appearance of the patient. The outcome of this step is a 3D estimate of the patient's torso. This patient specific 3D model has to be rigidly transformed in space such that its position coincides with the patient's location in the image. Additionally, the surgeon or the patient have the option to parametrically control the patient's 3D model. Modifications include variations to the size of the breasts, their position, shape etc. The modified torso is re-rendered onto the acquired imagery to provide a pre-visualization of the desired surgery. To be of practical use, the entire pipeline must operate on a mobile device in real time.
Landmark detection is a fundamental task in this pipeline, playing an important role in not only estimating the shape of the torso, but also to spatially track it. In this section, we describe our method to detect landmarks on the female torso. Nowadays, landmark detection is almost entirely solved with neural networks. Therefore, we also resort to a learning approach to address the problem of detecting landmarks on the female torso.

3.1. Network Design Principles

Neural networks are typically large, computation hungry models that require workstations with GPUs to run close to real time. Since our goal is to run landmark detection on mobile devices for surgical visualizations, we additionally need to take the following factors into account.

3.1.1. Accuracy

The landmarks detected on the torso serve the dual purpose of torso fitting and tracking. Naturally, for such surgical pre-visualizations, the accuracy of the fit is of primary importance. Several recent methods have shown that deeper networks tend to perform better than their shallow counterparts, given a sufficient training corpus. This however comes at the cost of increased inference time.

3.1.2. Inference Time

For the pre-visualization to be practically useful for the patient, landmarks must be detected faster than real time (> 30 FPS) in order to provide a leeway for subsequent processing and visualization. The inference time or the time taken to predict landmarks for a single frame on average is governed by two main factors; the availability of a GPU, and the size of the network. Given that high-end mobile devices such as the Ipad-Pro are already equipped with specialized hardware for neural network applications, it is reasonable to assume that in the near future, the hardware will not be a limiting factor for faster than real time landmark detection on almost all mobile platforms. Hence the size of the network will play the most important role. Unfortunately, as noted before, bigger networks perform better. Therefore, the speed-accuracy tradeoff has to be considered when designing architectures for mobile platforms.

3.1.3. Temporal Consistency

Almost as important as accuracy, is the temporal consistency of landmarks. Poor temporal stability of the landmarks can result in jittery pre-visualizations which reduce the reliability of the solution. To this end, we use a differentiable argmax operation to predict landmark positions directly as opposed to predicting heatmaps (please refer to section 3.3 for more details).

3.2. Stacked Hourglass with bottleneck residual blocks

The stacked hourglass network [1] has been proven to work in a number of landmark detection applications. It has been successfully applied to detect landmarks on the human body, hands and faces [1, 2, 4]. Therefore, it was a natural first choice for our application too.

The architecture of the hourglass is designed to leverage information from multiple scales of the input. The authors of the stacked hourglass observe that while local evidence is essential for identifying features like faces and hands, a final pose estimate requires a coherent understanding of the full body. The hourglass is a simple, encoder-decoder design with several skip connections that reuse information across corresponding scales of the encoder and decoder. The network accepts an input image of resolution 256 x 256 pixels and compresses it to a resolution of 4x4 pixels at the final stage of the encoder. Features are spatially reduced to lower resolutions using a combination of convolutional and max pooling layers. At each max pooling step, the network branches off and applies more convolutions at the original pre-pooled resolution. After reaching the lowest resolution, the network begins the top-down sequence of upsampling and performs a combination of features across scales. To bring together information across two adjacent resolutions, the hourglass architecture performs a nearest neighbor upsampling of the lower resolution followed by an element-wise addition of the two sets of features. Upon reaching the output resolution of the network, two consecutive rounds of 1x1 convolutions are applied to produce the final network predictions. The output of the network is a set of heatmaps where for a given heatmap the network predicts the probability of a landmark's presence at each and every pixel. Our architecture uses residual blocks for each layer of the hourglass. Similar to the original proposal, we use bottleneck residual blocks with filter sizes of 3x3 at every layer. This restricts the total number of parameters at each layer, and keeps overall memory usage in check. Once such an hourglass network with residual blocks has been constructed, we append another hourglass module to its end. This additional stack feeds the output of the first stage as an input to the next. This provides the network with a mechanism to redefine its features or make incremental improvements to the detected landmarks. This is made possible because the first hourglass stage also predicts its estimate of the landmark heatmaps which are supervised with ground truth.

In our application, we are not interested in estimating the pose of the whole body, and restrict ourselves to estimating the shape and pose of the torso. Nonetheless, it is a natural observation that a global context derived from the position of the patient's shoulders and arms will aid in localizing the rough position of the torso. Further, in the context of cosmetic breast surgery, the images captured by the surgeon are under controlled conditions where the patient is in a set of standard poses. Typical poses used for such captures include the frontal and profile views. We account for these priors when rendering our synthetic dataset as explained in section 4.

3.3. Differentiable Argmax

Most existing convolutional networks that regress heatmaps are trained with ground truth heatmaps. Ground truth heatmaps are generated by applying a spatial gaussian filter to the position of the landmarks. The standard deviation of this gaussian filter is manually specified and the position of all landmarks in the ground truth are blurred using the same. However, certain landmarks in the training set are localized with higher uncertainty due to the underlying feature. For example, in the case of the female torso, landmarks like the nipples, and areola are easier to unambiguously identify and therefore to annotate than those on the breast. Training networks with heatmaps created from isotropic Gaussian

kernels is enforcing the assumption that localization of all landmarks is equally (un)certain. During test time, the position of the landmarks are extracted from the predicted heatmaps using an argmax operation. While this is a simple way of extracting positions from the heatmaps, it is limited by its ability to predict only integer positions. The operation is also not differentiable, meaning that the network cannot directly regress landmark positions, while remaining fully convolutional.
Unlike previous methods in landmark detection, we choose to represent the output of our convolutional networks as latent heatmaps without ground truth supervision. This provides the hourglass with the flexibility to be more confident about certain landmarks than others and to represent them using anisotropic non-gaussian distributions.The latent heatmap output by the global hourglass is passed through a channel-wise spatial softmax to ensure that each channel is a probability distribution over the landmark's position in the image. Then, we perform a softargmax operation on the landmark heatmaps to extract landmark positions as a batch size x number of landmarks x 2 vector. Since the soft-argmax operation boils down to a weighted average, it is fully differentiable unlike the argmax. Extracting landmark positions this way enables us to train the hourglass modules using only ground truth landmark positions without having to create ground truth heatmaps, while at the same time ensuring that the landmark positions are represented inside the network as a heatmap, and therefore keeping the network fully convolutional. Our full architecture is shown in figure 2.

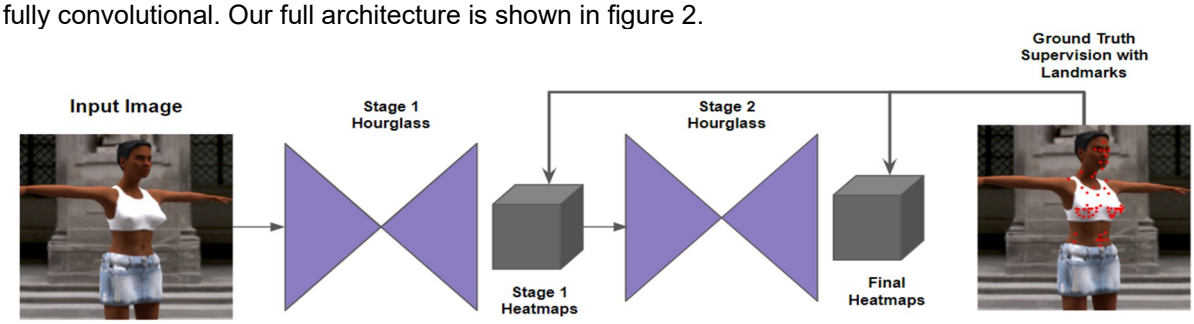

Figure 2: Our two stage hourglass network where the heatmaps are supervised using 2D ground landmarks thanks to the softargmax operation.

4. Datasets

For a data-driven application such as ours, the availability of a well annotated dataset is of utmost importance. Unfortunately, to the best of our knowledge, there exists no publicly available dataset of female torsos in the context of landmark detection for surgical visualizations. Building such a dataset is primarily limited due to privacy concerns and is further burdened by the cost of annotating data. Therefore we resort to generating photo-realistic synthetic data and learning from such images. In the following section, we describe our synthetic dataset creation pipeline.
Using an openly available parametric model for human character generation, we procedurally generated 600 detailed digital humans. Each character was created with a random hairstyle, a random body pose and random clothing. To consider the wide variation in the shape of female torsos, parameters of the breasts were also altered to create a variety of plausible shapes. On these characters, we define the landmarks shown in figure 3.

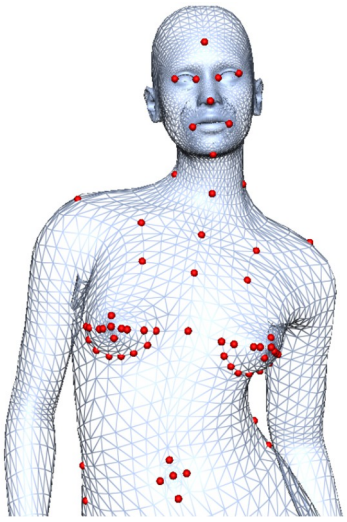

Figure 3. Landmarks defined on a template female torso

These characters were then rendered in different viewpoints and environments by path tracing. During rendering, skin reflectance properties like sub-surface scattering were also considered to result in realistic renderings of these characters. Some example renderings from our synthetic dataset are shown in figure 4.

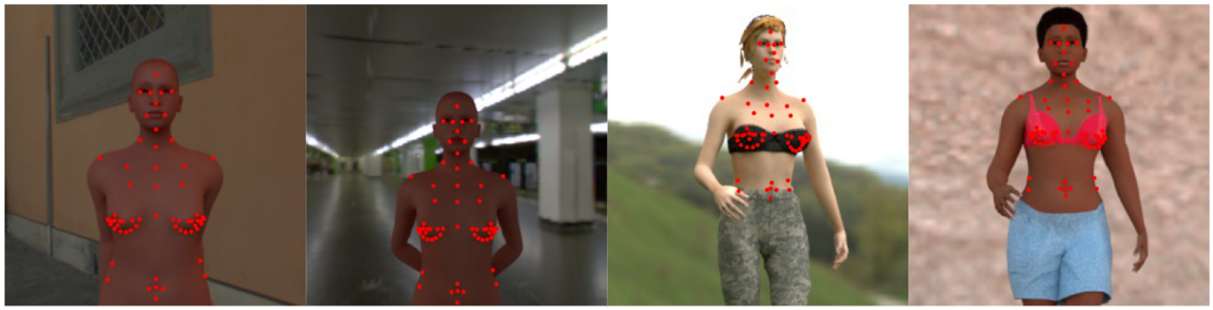

Figure 4: Examples of synthesized characters from our database with landmark annotations. To generalize to diverse capture scenarios, we generate both naked and clothed characters. Notice the variation in clothing, pose and shape of these characters.

In total, we created a database of 22,760 samples containing 600 characters, and 64 landmarks. We refer to this dataset as *SynHuman-Train*.In addition to *SynHuman-Train*, we additionally created 100 new characters, with novel shape and clothing for validation. These characters were rendered in novel environments and viewpoints in the same way as *SynHuman-Train*. Our validation dataset consisted of 10,589 samples, spanning 100 characters, and ground truth landmarks. We refer to this dataset as *SynHuman-Val*.

5. Implementation Details

We use an L2 loss to supervise the output of the network with the ground truth landmarks. We implemented our solution in Pytorch and used a single Nvidia 1070 GPU to train our model. We used the Adam optimizer and a batch size of 8.

6. Results

In this section we show quantitative and qualitative results that emphasize the performance of our application. All results are reported on the *SynHuman-Val* dataset.

6.1. Quantitative Results

We use the Percentage Correct Keypoints (PCK) [1] metric for quantitative evaluation. In figure 5, we plot the PCK metric evaluated at different thresholds for the *SynHuman-Val* dataset. We obtain an AUC of 0.94 which is comparable to state of the art landmark detectors for faces and human bodies.

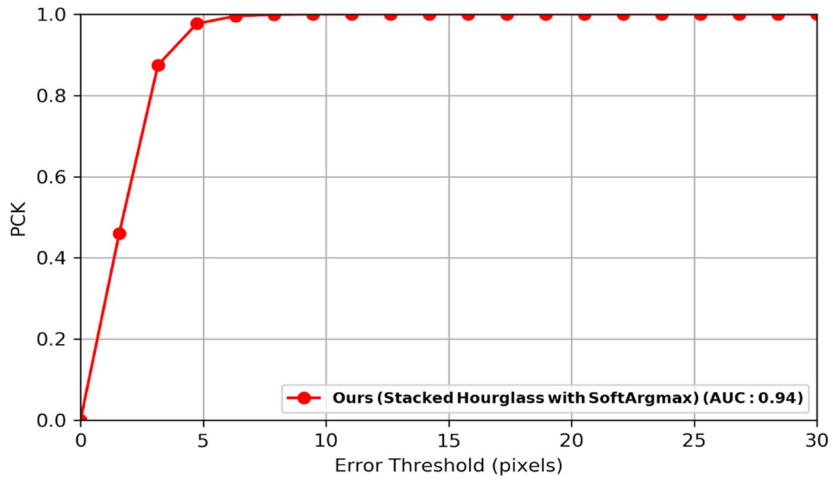

Figure 5: PCK plot for our method on the SynHuman-Val dataset.

6.2. Qualitative Results

In figure 6, we plot predicted landmarks on a few samples from the *SynHuman-Val* dataset. We also evaluated our method with real imagery captured from hospitals and found our method to generalize remarkably well. However, for privacy reasons, we do not display the results on the real imagery here.

Figure 6: Examples of predictions on the SynHuman-Val dataset.

7. Conclusion

In this paper we presented ARSynth, a robust pipeline to track human upper bodies, based only on realistically generated synthetic datasets and trained utilizing adaptation of state-of-the-art deep neural networks. We demonstrated the capability of such networks to generalize well on unconstrained real world images.

What we did not demonstrate in this paper is how the 3D upper torso fitting is performed, which combined with the tracker and a proper appearance modeling can be utilized for Augmented reality applications. We leave this for future work.

A direct application to this method is the live pre-visualization of breast surgical operations in an Augmented Reality setting. The method however is general and can be easily applied to other body parts. More importantly, these methods span other fields such as 'Body Technology for Apparel', where a virtual cloth fitting from home use application can be envisaged.

References

[1] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2025–2033, July 2017

[2] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz. Hand pose estimation via latent 2.5d heatmap regression. Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI, pages 125–143, 2018.

[3] K. He, G. Gkioxari, P. Doll´ar, and R. B. Girshick. Mask RCNN. IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 2980–2988, 2017.

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 1021–1030, 2017.

[5] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affinity fields. CoRR, abs/1611.08050, 2016

[6] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. P. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4177–4187, 2016.

[7] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. CoRR, abs/1603.01249, 2016